Award Number:
W81XWH-07-2-0112

TITLE:
Literature Mining of Pathogenesis-Related Proteins in Human Pathogens for Database Annotation

PRINCIPAL INVESTIGATOR:
Cathy H. Wu, Ph.D.
Zhang-Zhi Hu, M.D.

CONTRACTING ORGANIZATION:
Georgetown University Medical Center
Washington, DC 20007

REPORT DATE:
October 2008

TYPE OF
REPORT: *Annual*

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
19-10-2008	Annual Report	20 Sep 2007-19 Sep 2008

4. TITLE AND SUBTITLE

Literature Mining of Pathogenesis-Related Proteins in Human Pathogens for Database Annotation

5a. CONTRACT NUMBER

W81XWH-07-2-0112

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Wu, Cathy H., Ph.D.
Hu, Zhang-Zhi, M.D. Email: wuc@georgetown.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Georgetown University Medical Center
3300 Whitehaven Street NW, Suite 1200
Washington, DC 20007

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Army Medical Research and Material Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Biomedical literature represents the primary source of experimental data and biological knowledge. This project aims to develop a text mining system for pathogens of biodefense relevance, focusing on mining pathogen-host protein-protein interactions (PH-PPI). We developed a Support Vector Machine (SVM)-based system to identify abstracts containing PH-PPI information using an annotated corpus of 1360 MEDLINE abstracts as the training set. It achieved good performance on document classification with a precision of over 80% among top 50 ranked abstracts. The SVM-based method is further augmented with other text mining tools (such as PIE) for mining and tagging PPI information. As part of an effort in enabling text mining tools for real-world applications, we are developing a basic framework, iProLINK, to connect text mining tools with ontology and systems biology for the biomedical research community. The PH-PPI text mining system developed in the first year will be coupled with the iProXpress proteomic data analysis system into a "Pathogen Mining System" for the analysis of pathogen proteomics data.

15. SUBJECT TERMS

Text mining, pathogenesis, pathogen-host protein-protein interaction, literature corpus, machine learning, support vector machine (SVM), proteomics data analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USARMMC
U	U	U	UU	36	19b. TELEPHONE NUMBER *(Include area code)*

Reset

Table of Contents

INTRODUCTION:

Due to the heightened concern about bioterrorism and emerging/reemerging infectious diseases, there are growing interests and pressing needs in speeding up the basic research as well as data mining of pathogenesis-related proteins in pathogens of military relevance, which may lead to better targets for disease diagnosis, prevention and therapy. This project specifically focuses on pathogenesis-related protein data mining from scientific literature by developing an automated text mining system to facilitate literature-based curation of such proteins (1st year), and from proteomics and functional genomics data through an integrated protein bioinformatics analysis system (2nd year, revised). We refer to the project as Pathogen Mining System (https://pir5.georgetown.edu/wiki/TATRC). The text mining system development primarily concerns the pathogen-host protein-protein interaction (PH-PPI) information from MEDLINE abstracts, and the proteomics and genomics data mining concerns the analysis of proteomics data from Burkholderia under simulated growth condition, and of transcriptional regulatory pathway data from Ebola virus-infected macrophage.

BODY:

The primary objective of the first year of the project is to develop a text-mining system to identify pathogenesis-related papers and extract information on pathogenicity and host-pathogen interactions. There are three tasks:

- ***Task1 (M01-03): Compilation of training and benchmarking literature corpus.*** Manual compilation of literature corpus as a positive training set of 300 pathogenesis-related papers with pathogen-host protein-protein interaction information.
- ***Task2 (M04-09): Development and evaluation of text-mining algorithms.*** Development of a text-mining system for document retrieval, entity recognition, and document categorization. Named-entity tagging tools as well as algorithms for document classification and information extraction, including machine learning and rule-based methods will be evaluated.
- ***Task3 (M10-12): Development of web interface for automated literature mining.*** Development of web-based graphical user interface for query submission and for literature mining result display with automatically tagged abstracts.

I. Literature data sets for machine learning algorithm training

Literature data sets (literature corpus) consisting of positive and negative data are necessary for training machine learning algorithms, such as Supporter Vector Machine (SVM), for text mining of pathogenesis-related pathogen and host proteins from literature. We focused on specific pathogen and host protein-protein interactions (PH-PPI). Unlike those for protein-protein interactions of the same species taking place within an organism, curated positive training data sets are rare for PH-PPI, especially for bacterial PH-PPI, and most such data are buried in the literature. Also because the bacterial PH-PPI information is much more difficult to distinguish from the same-species PPI than viral PH-PPI information would, we decided to separate training set for the bacterial PH-PPI from that of viral PH-PPI, and to concentrate on the former. Thus,

we generated the literature training sets through manual curation of a set of ~2000 abstracts retrieved from PubMed based on query terms "bacterial pathogen and protein interaction".

1. Positive literature set of PH-PPIs. We compiled 300 abstracts (PMIDs) that are reviewed to contain PH-PPI, and the sentences providing the evidence for such interactions are also tagged (highlighted). The sources for deriving the set of literature also include protein databases (UniProtKB and IntAct) where literature with protein interactions is cited for protein entries. Of the 300 abstracts, ~54% are for viral-host PPI, which are all derived from literature cited in databases; while ~46% are for bacterial-host PPI, most of which are from PubMed search. Because the primary interests of pathogens for the USAMRIID are on CDC category A/B viral and bacterial pathogens, the abstracts for training have a balanced coverage of the bacterial and viral groups of organisms. In the training set, viral pathogens include Ebola, Lassa, HIV, HBV, and bacterial pathogens include Yersinia pestis, Bacillus anthracis, Salmonella, and Shigella. In most cases the host is human, but may also include other mammal species.

2. Negative literature set of PH-PPIs. Of the ~2000 abstracts retrieved from PubMed based on general keyword search "bacterial host protein interaction", ~1225 abstracts were manually selected as negative ones, which may describe pathogen gene- or protein-related information but clearly lack of specific PH-PPI information.

The data sets for bacterial PH-PPI are available at http://pir.georgetown.edu/staff/huz/tatrc/ (*tatrc_dataset_positive.html* and *tatrc_dataset_negative.html*), including 135 positive and 1225 negative abstracts. Evidence sentences in the positive abstracts were also annotated. The data set is currently for internal use and will eventually be made public for use in developing text mining algorithms by the text mining community.

II. Machine learning algorithm development for text mining of pathogenesis proteins

We developed and evaluated machine learning-based text-mining methods for retrieving MEDLINE abstracts containing pathogen and host protein-protein interaction information based on the literature training set. We used a publicly available Support Vector Machine (SVM) package, SVM^{light} (see http://svmlight.joachims.org/), to train the classifier, and tested and evaluated both abstract- and sentence-based classifiers to recognize PH-PPI-containing abstracts. Detailed methodology and results are described in a research paper (Xu *et al*., 2008) to be presented at the 2008 IEEE conference on Bioinformatics and Biomedicine (BIBM 2008) (http://www.ischool.drexel.edu/ieeebibm/bibm08/).

1. Abstract-based algorithm. The training task can be at abstract level (ALT) to build a system to rank a set of abstracts. The abstracts in the dataset were preprocessed first by normalizing the nouns, verbs, and adjectives, followed by extracting the unigrams and bigrams in both title and abstract to construct the sample features. The SVM was trained to classify these 1360 abstracts (both positive and negative) by 10-fold cross-validation. Given a threshold value, abstracts with scores higher than the threshold from the classifier were assigned positive, while those with lower scores labeled negative. The classification was based on the total feature of the abstract. We tried different kernel functions in SVM including linear function, polynomial, and RBF and found linear function was the best.

2. Sentence-based algorithm. The training task can also be at sentence level (SLT) to build a system to rank the abstracts. Individual sentences from abstracts were first extracted and labeled with corresponding PubMed ID (PMID) appended with a sequential number of the sentence in the given abstract. The sentences were then preprocessed similarly as above in the abstract-based algorithm. Untagged sentences from positive abstracts were not used for training but included in the test dataset only. The SVM was trained with linear function at the sentence-based, and 10-fold cross-validation was used to construct training and test dataset. Each sentence received a score from the classifier, and the highest sentence score would be assigned to the abstract as the final discriminating value. Similar to ALT method, a threshold value was set to assign positive or negative abstracts from the classifier, but the classification in SLT method was based on the feature of sentences.

3. Results and comparison between ALT and SLT methods. The testing results of the trained SVM were evaluated using the ROC curve depicting the relationship between the true positive (TP) and false positive (FP) rates (Figure 1). In the high specificity area (specificity=1-FP, towards the left of the ROC curve), given the same sensitivity (TP), the sentence-based method gave higher specificity (red-line) than the abstract-based (blue-line); while in the high sensitivity area (sensitivity=TP, towards the top of the ROC curve), the two methods seemed to have little difference. For example, the top 200-scored abstracts from the classifier using sentence-based method contained 61% true positive abstracts, compared to 53% with abstract-based method. The results suggest that the sentence-based training method tends to have better performance than the abstract-based method for retrieving pathogen host PPI abstracts. We also extended the SVM training to feature selection to enhance its performance.

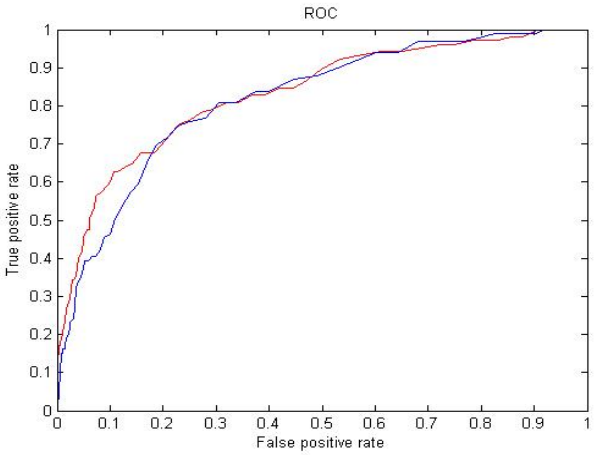


Figure 1. Receiver operating characteristics curve (ROC) analysis of ALT (blue) and SLT (red).

4. Feature selection method and information gain. We investigated the inclusion of a feature selection method (i.e., information gain) into the machine leaning system. We compared *no feature selection* method with *Information Gain* feature selection on both abstract and sentence levels. We found that *Information Gain* reduced the dimension of Vector Space and could improve the performance of the SVM than *no feature selection*. Moreover, the results showed that the sentence-level SVM (training based on highlighted sentences) had better performance and greater prospect than the abstract-based method.

III. Evaluation of existing text mining tools on the PH-PPI data sets

While developing and evaluating the SVM-based text mining system for PH-PPI during the first year of the project, we are also exploring the existing text mining tools that can be useful for text

mining of PH-PPI information. These public text mining tools include PIE (Kim et al., 2008), iHOP (Fernández et al., 2007), and others as included in MetaServer (Leitner et al., 2008), which is a central sever integrating text mining tools participating in the BioCreative Challenge Evaluation for molecular (gene and protein) data from literature (Hirschman et al., 2005). Protein-protein interaction text mining has been a major task in the 2nd BioCreative Challenge Evaluation (Wilbur et al., 2007).

We evaluated the PPI text mining tool PIE (Protein Interaction information Extraction, http://pie.snu.ac.kr/index.php) using the curated positive data set for bacterial as well as the viral PH-PPI. PIE highlights sentences in abstracts that contain protein interaction information, in which the detected words/phrases for the interacting proteins and the interaction relations are also distinguished. Table 1 (bacterial set) and 2 (viral set) summarize the comparison of the PIE PPI extraction with the manual annotated abstracts and sentences.

Table 1. Comparison of PIE text mining of PPI to the manual bacterial data set

		# Abstracts	% Data set
Abstract level	Manually-tagged bacteria data set	135	100%
	Positive abstracts tagged by PIE	110	**81.5%**
	Positive abstracts not tagged by PIE	25	18.5%
	Abstracts with >=1 manually-identified sentence tagged by PIE	70	51.9%
	Abstracts with no manually-identified sentence tagged by PIE	65	48.1%
Sentence level	Manually-tagged (positive) sentences in data set	247	100%
	Positive sentences tagged by PIE	98	39.7%
	Positive sentences missed by PIE	149	60.3%
	Sentences tagged by PIE in data set	298	100%
	Positive sentences tagged by PIE	98	32.9%
	Negative sentences tagged by PIE	200	67.1%

Table 2. Comparison of PIE text mining of PPI to the manual viral data set

		# Abstracts	% Data set
Abstract level	Manually-tagged virus data set	170	100%
	Positive abstracts tagged by PIE	163	**95.9%**
	Positive abstracts not tagged by PIE	7	4.1%
	Abstracts with >=1 manually-identified sentence tagged by PIE	145	85.3%
	Abstracts with no manually-identified sentence tagged by PIE	25	14.7%
Sentence level	Manually-tagged sentences (positive) in the data set	279	100%
	Positive sentences tagged by PIE	205	**73.5%**
	Positive sentences missed by PIE	74	26.5%

The results show that PIE recognizes ~82% of the manually tagged abstracts and ~40% manually tagged sentences for the bacterial data set, and recognizes ~96% manually tagged abstracts and 74% manually tagged sentences for the viral data set. While we need to compare the PIE's performance with other similar tools on the same data set, the relatively high recognition of positive abstracts by PIE is a desired feature for retrieving the PH-PPI containing abstracts to

facilitate the manual curation efforts. Therefore the PIE tool can augment the pathogen mining system for this project. The detailed evaluation results of the PIE tool are available at: http://pir.georgetown.edu/staff/huz/tatrc/dataset/ with the bacterial set (PIE_evaluation_bacterial_positive.mht) and the viral set (PIE_evaluation_viral_positive.mht).

IV. iProLINK framework to link text mining to ontology and systems biology

Another ongoing effort relevant to the project on the PH-PPI text mining is the iProLINK framework development, an effort in bringing together text mining, biological ontology and systems biology communities to develop text mining tools that can be broadly utilized by the biology communities for real-world applications.

The ever-increasing scientific literature and the exponential growth of large-scale molecular data have prompted active research in biological text mining to facilitate literature-based curation of molecular databases. Meanwhile, systems biology and bio-ontologies are emerging as critical tools in biological research where complex data in disparate resources are generated, integrated and analyzed. Both rely on literature for data annotation and analysis. The challenges facing us are to develop broadly utilized text mining tools and systems that need to involve both developers and users for system development and evaluation. iProLINK, extending from a previously developed text mining resource (Hu et al., 2004), is designed as a framework for linking text mining tools with ontology and systems biology. The framework focuses on text mining of protein-protein interaction, including the protein posttranslational modification such as phosphorylation, which can be applied to curation of molecular and ontological data and analysis of systems biology data.

The framework consists of two major components: a user interface for text mining of PPI from an integrated tool server and software modules to allow text mining outputs to be created, ranked, and used by the community. Use cases are presented for assessing the gaps and making recommendations for future development. The detailed components and case studies are described in a research paper (Hu *et al*., 2008) to be presented at the 2008 IEEE conference on Bioinformatics and Biomedicine (http://www.ischool.drexel.edu/ieeebibm/bibm08/).

The iProLINK framework will benefit the current Pathogen Mining project by not only maximally utilizing the different tools developed by the text mining community and providing an interface for community access, but also encouraging the use and application of these tools in the real-world applications such as assisting genomic and proteomic data analysis and pathogen data mining. As will be mentioned below (section V), the second year of the Pathogen Mining project has been revised to focus on the collaborative work with the USAMRIID on bacterial pathogen proteomics data analysis using the iProXpress system developed at PIR (Huang et al, 2007).

V. Collaboration with USAMRIID research groups

In the beginning of the project, we met with the USAMRIID research groups and discussed the research activities in their labs complimentary to the current project. We identified areas that were of great importance and high priority for the USAMRIID. One of the projects was the legacy proteomics data for Burkholderia from Dr. Brad Powell's lab that need to be re-annotated

and reanalyzed using the iProXpress system. It has been approved that the second year for the project will focus on the bacterial pathogen proteomics data analysis. The CRADA has been approved by both Georgetown University and USAMRIID for Dr. Powell's data.

The prior bacterial proteomics data for Burkholderia (>5 years old) needs to be reanalyzed due to the continuous updates to the relevant bacterial protein databases and/or annotations, as well as accumulation of literature information regarding prior unknown genes. The objective of this collaboration is to use the integrated proteomics analysis system, iProXpress, coupled with the current TATRC-funded project, Pathogen Mining System, to facilitate the re-evaluation and functional interpretation and hypothesis formulation from the legacy data.

KEY RESEARCH ACCOMPLISHMENTS:

- We manually curated pathogen-host PPI literature data sets that are necessary for the machine learning method as well as beneficial to the text mining community when becoming publicly available.
- We developed and evaluated the SVM methods for classifying the abstracts with PH-PPI information, whose overall performance is best when using sentence level training and feature selection.
- We identified and evaluated existing public text mining tools such as PIE that can be augmenting the Pathogen Mining System.
- We initiated a community collaborative effort under the iProLINK framework, which will be of great benefit to the Pathogen Mining System.
- We established close collaborations with USAMRIID research groups to analyze pathogen genomic and proteomic data that will take advantage of the PH-PPI text mining.

REPORTABLE OUTCOMES:

1. Cooperative Research And Development Agreement between Georgetown University and USAMRIID
2. Two research papers were generated from the project, one reporting the SVM-based PH-PPI text mining system (Xu et al., 2008), the other reporting an integrated text mining framework for text mining and biology communities (Hu et al., 2008).
3. A workshop presentation for the 2009 PAG XVI (Plant and Animal Genome Conference) on the iProLINK framework (Hu and Hirschman, 2009).

CONCLUSIONS:

Biomedical literature represents the primary source of experimental data, and developing text mining systems for mining such data for pathogens of biodefense relevance is the main objective for the first year of the project. We focus on text mining of the host-pathogen protein-protein interactions. We developed an SVM-based automated system to identify MEDLINE abstracts containing HP-PPI information. We observed that feature selection was effective not only in reducing the dimensionality of features to build a compact system, but also in improving

document classification performance. We also observed abstract-level systems and sentence-level systems yielded different classification of MEDLINE abstracts, and the combination of these systems could improve the overall document classification.

To augment the SVM-based PH-PPI mining methods, we also explored the public text mining tools for the PH-PPI mining. We performed preliminary evaluation on the PPI extraction tool PIE, and the results showed encouraging performance at least at the abstract level, suggesting that PIE can be potentially integrated into the Pathogen Mining System for improving the overall text mining capabilities of the system. Exploring public text mining tools is also part of the initiative by PIR in order to develop a basic framework to bring together the text mining and biological communities to better develop text mining tools for real-world applications.

Our second year tasks have been revised to focus on the integrated analysis of the pathogen proteomics data, which will be done in a coordinated fashion to the development of iProLINK framework, which will facilitate the use of text mining results to the annotation and analysis of systems biology data, including genomic and proteomic data for pathogens of biodefense and military relevance.

REFERENCES:

1. J.M. Fernández, R. Hoffmann, A. Valencia. iHOP web services. Nucleic Acids Res. 35:W21-6, 2007.
2. L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, Overview of BioCreative: Critical Assessment of Information Extraction for Biology, BMC Bioinformatics 6(Suppl 1):S1, 2005.
3. Z.Z. Hu, K.B. Cohen, L. Hirschman, A. Valencia, H. Liu, M.G. Giglio, C.H. Wu. iProLINK: A Framework for Linking Text Mining with Ontology and Systems Biology. In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2008), Philadelphia, 2008, *in press*
4. Z.Z. Hu and L. Hirschman. Linking Text Mining with Ontology and Systems Biology for Database Curation. (Workshop abstract) Plant and Animal Genome Conference (PAG) XVII, San Diego, CA, January 10-14, 2009
5. Z.Z. Hu, I. Mani, V. Hermoso, H. Liu and C.H. Wu. iProLINK: an integrated protein resource for literature mining. Comput Biol Chem 28: 409-416, 2004.
6. H. Huang, Z.Z. Hu, C.N. Arighi, C.H. Wu. Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. Front Biosci. 12:5071-88, 2007.
7. S. Kim, S.Y. Shin, I.H. Lee, S.J. Kim, R. Sriram, B.T. Zhang. PIE: an online prediction system for protein-protein interactions from text. Nucleic Acids Res. 36(Web Server issue):W411-5, 2008.
8. F. Leitner, M. Krallinger, C. Rodriguez-Penagos, J.Hakenberg, C Plake et al. Introducing meta-services for biomedical information extraction. Genome Biology, 9(Suppl 2):S6, 2008.
9. J. Wilbur, L. Smith and L. Tanabe. BioCreative 2. Gene Mention Task, In Proceedings of the Second BioCreative Challenge Evaluation Workshop, 2007, pp. 7–16.
10. G. Xu, L. Yin, M. Torii, Z. Niu, C. Wu, Z.Z. Hu, H. Liu. Document Classification for Mining Host Pathogen Protein-Protein Interactions. In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2008), Philadelphia, 2008, *in press*

APPENDICES:

- Cooperative Research And Development Agreement
    - Title :Reanalysis and Functional Interpretation of Proteomics Data from Bacterial Cells under Simulated Growth Condition
    - Between Georgetown University and USAMRIID (US Army Medical Research Institute of Infectious Disease)
    - USAMRMC Control No: W81XWH-09-0003
- The PH-PPI text mining paper (BIBM2008) (Xu et al., 2008) (in press)
    - Title: Document Classification for Mining Host Pathogen Protein-Protein Interactions
- The iProLINK text mining framework paper (BIBM2008) (Hu et al., 2008) (in press)
    - Title: A Framework for Linking Text Mining with Ontology and Systems Biology
- The PAG workshop abstract (Hu and Hirschman, 2009)
    - Title: Linking Text Mining with Ontology and Systems Biology for Database Curation

COVER SHEET
COOPERATIVE RESEARCH AND DEVELOPMENT AGREEMENT

[NOTE: This Cover Sheet is for internal management purposes only. It is not part of the Agreement and neither party is bound to anything contained in it]

Title: Reanalysis and Functional Interpretation of Proteomics Data from Bacterial Cells under Simulated Growth Condition

Effective Date : 10-07-2008
Expiration Date: 10-07-2010

USAMRMC Control No. W81XWH-09-0003
DA/TTPO Control No.

Primary NTIS Subject Code/Title: 57K
Secondary NTIS Subject Code/Title:
STO Code/Title:
Oracle #: 92951
Concurrence obtained from appropriate RAD/USAMMDA/CBMS-JPMO program managers: YES/RAD 4

Laboratory:
USAMRIID
MCMR-UIZ-D
1425 Porter Street
Fort Detrick, MD 21702-5011
Voice Phone: 301-619-6886 FAX Phone: 301-619-8379

Lab's Technical POC
Dr. Bradford Powell
USAMRIID / MCMR-UIB
1425 Porter Street, Fort Detrick, MD 21702-5011
Voice Phone: 301-619-4933 FAX Phone: 301-619-2152
Email: Bradford.powell@amedd.army.mil

Lab's Legal Counsel:
Commander, U.S. Army Medical Research and Materiel Command
ATTN: MCMR-JA (Mr. Robert L. Charles)
Fort Detrick, Frederick, MD 21701-5012
Voice Phone: 301-619-2065 FAX Phone: 301-619-5034

Cooperator's POCs:
Dr. Cathy H. Wu (Scientific POC)
Georgetown University Medical Center
3300 Whitehaven Street, NW; Suite 1200
Washington, DC 20007
Phone: 202-687-1039 Fax: 202-687-0057
Email: wuc@georgetown.edu

Silvana T. Alcocer (Administrative)
IP & Contract Administrator; Office of Technology Commercialization
Georgetown University
3300 Whitehaven Street, NW; Suite 1500
Washington, DC 20007
Phone: 202-687-0843 Fax: 202-687-3111
Email: alcocers@georgetown.edu

Summary: In this collaboration, USAMRIID will provide MS data comprising protein lists and other information of relevance for matched data sets to be re-analyzed by the Pathogen Mining system.

Bob Charles reviewed 9/15/2008.

A COOPERATIVE RESEARCH AND DEVELOPMENT AGREEMENT

Between

Georgetown University
37th and O Streets, NW
Washington, District of Columbia 20057
(Cooperator)
and
U.S. Army Medical Research Institute of Infectious Diseases
Fort Detrick, Maryland 21702-5011
(Laboratory)

Article 1. Background

1.00 This Agreement is entered into under the authority of the Federal Technology Transfer Act of 1986, 15 U.S.C. 3710a, et seq., between the Cooperator and the Laboratory, the parties to this Agreement.

1.01 Laboratory, on behalf of the U.S. Government, and Cooperator desire to cooperate in research and development on Reanalysis and Functional Interpretation of Proteomics Data from Bacterial Cells under Simulated Growth Condition according to the attached Statement of Work (SOW) described in Appendix A. NOW, THEREFORE, the parties agree as follows:

Article 2. Definitions

2.00 The following terms are defined for this Agreement as follows:

2.01 "Agreement" means this cooperative research and development agreement.

2.02 "Invention" and "Made" have the meanings set forth in Title 15 U.S.C. Section 3703(9) and (10).

2.03 "Proprietary Information" means information marked with a proprietary legend which embodies trade secrets developed at private expense or which is confidential business or financial information, provided that such information:

(i) is not generally known, or which becomes generally known or available during the period of this Agreement from other sources without obligations concerning their confidentiality;

(ii) has not been made available by the owners to others without obligation concerning its confidentiality; and

(iii) is not already available to the receiving party without obligation concerning its confidentiality.

(iv) is not independently developed by or on behalf of the receiving party, without reliance on the information received hereunder.

2.04 "Subject Data" means all recorded information first produced in the performance of this Agreement.

2.05 "Subject Invention" means any Invention Made as a consequence of, or in relation to, the performance of work under this Agreement.

Article 3. Research Scope and Administration

3.00 Statement of Work. Research performed under this Agreement shall be performed in accordance with the SOW incorporated as a part of this Agreement at Appendix A. It is agreed that any descriptions, statements, or specifications in the SOW shall be interpreted as goals and objectives of the services to be provided under this Agreement and not requirements or warranties. Laboratory and Cooperator will endeavor to achieve the goals and objectives of such services; however, each party acknowledges that such goals and objectives, or any anticipated schedule of performance, may not be achieved.

3.01 Review of Work. Periodic conferences shall be held between the parties for the purpose of reviewing the progress of work. It is understood that the nature of this research is such that completion within the period of performance specified, or within the limits of financial support allocated, cannot be guaranteed. Accordingly, all research will be performed in good faith.

3.02 Principal Investigator. Any work required by the Laboratory under the SOW will be performed under the supervision of Dr. Bradford Powell, U.S. Army Medical Research Institute of Infectious Diseases (USAMRIID) 1425 Porter Street, Fort Detrick, MD 21702-5011, Phone: 301-619-4933, Fax 301-619-2152 and Email: bradford.powell@amedd.army.mil, who, as co-principal investigator has responsibility for the scientific and technical conduct of this project on behalf of the Laboratory. Any work required by the Cooperator under the SOW will be performed under the supervision of Dr. Cathy H. Wu, Georgetown University Medical Center, 3300 Whitehaven Street, NW, Suite 1200, Washington, DC 20007, Phone: 202-687-1039, Fax: 202-687-0057, and Email: wuc@georgetown.edu, who, as co-principal investigator has responsibility for the scientific and technical conduct of this project on behalf of the Cooperator.

3.03 Collaboration Changes. If at any time the co-principal investigators determine that the research data dictates a substantial change in the direction of

the work, the parties shall make a good faith effort to agree on any necessary change to the SOW and make the change by written notice to the addresses listed in section 12.05 Notices.

3.04 Final Report. The parties shall prepare a final report of the results of this project within six months after completing the SOW.

Article 4. Ownership and Use of Physical Property

4.01 Ownership of Materials or Equipment. All materials or equipment developed or acquired under this Agreement by the parties shall be the property of the party which developed or acquired the property, except that government equipment provided by Laboratory (1) which through mixed funding or mixed development must be integrated into a larger system, or (2) which through normal use at the termination of the Agreement has a salvage value that is less than the return shipping costs, shall become the property of Cooperator.

4.02 Use of Provided Materials. Both parties agree that any materials relating to them which were provided by one party to the other party will be used for research purposes only. The materials shall not be sold, offered for sale, used for commercial purposes, or be furnished to any other party without advance written approval from the Provider's official signing this Agreement or from another official to whom the authority has been delegated, and any use or furnishing of material shall be subject to the restrictions and obligations imposed by this Agreement.

Article 5. Patent Rights

5.00 Reporting. The parties shall promptly report to each other all Subject Inventions reported to either party by its employees. All Subject Inventions Made during the performance of this Agreement shall be listed in the Final Report required by this Agreement.

5.01 Cooperator Employee Inventions. Laboratory waives any ownership rights the U.S. Government may have in Subject Inventions Made by Cooperator employees and agrees that Cooperator shall have the option to retain title in Subject Inventions Made by Cooperator employees. Cooperator shall notify Laboratory promptly upon making this election and agrees to timely file patent applications on Cooperator's Subject Invention at its own expense. Cooperator agrees to grant to the U.S. Government on Cooperator's Subject Inventions a nonexclusive, nontransferable, irrevocable, paid-up license in the patents covering a Subject Invention, to practice or have practiced, throughout the world by, or on behalf of the U.S. Government. The nonexclusive license shall be evidenced by a confirmatory license agreement prepared by Cooperator in a form satisfactory to Laboratory.

5.02 Laboratory Employee Inventions. Laboratory shall have the initial option to retain title to, and file patent application on, each Subject Invention Made by its employees. The Laboratory agrees to grant an exclusive license to any invention arising under this Agreement to which it has ownership to the Cooperator in accordance with Title 15 U.S. Code Section 3710a, on terms negotiated in good faith. Any invention arising under this Agreement is subject to the retention by the U.S. Government of nonexclusive, nontransferable, irrevocable, paid-up license to practice, or have practiced, the invention throughout the world by or on behalf of the U.S. Government.

5.03 Joint Inventions. Any Subject Invention patentable under U.S. patent law which is Made jointly by Laboratory employees and Cooperator employees under the Scope of Work of this Agreement shall be jointly owned by the parties. The parties shall discuss together a filing strategy and filing expenses related to the filing of the patent covering the Subject Invention. If a party decides not to retain its ownership rights to a jointly owned Subject Invention, it shall offer to assign such rights to the other party, pursuant to Paragraph 5.05, below.

5.04 Government Contractor Inventions. In accordance with 37 Code of Federal Regulations 401.14, if one of Laboratory's Contractors conceives an invention while performing services at Laboratory to fulfill Laboratory's obligations under this Agreement, Laboratory may require the Contractor to negotiate a separate agreement with Cooperator regarding allocation of rights to any Subject Invention the Contractor makes, solely or jointly, under this Agreement. The separate agreement (i.e., between the Cooperator and the Contractor) shall be negotiated prior to the Contractor undertaking work under this Agreement or, with the Laboratory's permission, upon the identification of a Subject Invention. In the absence of such a separate agreement, the Contractor agrees to grant the Cooperator an option for a license in Contractor's inventions of the same scope and terms set forth in this Agreement for inventions made by Laboratory employees.

5.05 Filing of Patent Applications. The party having the right to retain title to, and file patent applications on, a specific Subject Invention may elect not to file patent applications, provided it so advises the other party within 90 days from the date it reports the Subject Invention to the other party. Thereafter, the other party may elect to file patent applications on the Subject Invention and the party initially reporting the Subject Invention agrees to assign its ownership interest in the Subject Invention to the other party.

5.06 Patent Expenses. The expenses attendant to the filing of patent applications shall be borne by the party filing the patent application. Each party shall provide the other party with copies of the patent applications it files on any Subject Invention, along with the power to inspect and make copies of all documents retained in the official patent application files by the applicable patent

office. The parties agree to reasonably cooperate with each other in the preparation and filing of patent applications resulting from this Agreement.

Article 6. Exclusive License

6.00 Grant. The Laboratory agrees to grant to the Cooperator an exclusive license in each U.S. patent application, and patents issued thereon, covering a Subject Invention, which is filed by the Laboratory subject to the reservation of a nonexclusive, nontransferable, irrevocable, paid-up license to practice and have practiced the Subject Invention on behalf of the United States.

6.01 Exclusive License Terms. The Cooperator shall elect or decline to exercise its right to acquire an exclusive license to any Subject Invention within six months of being informed by the Laboratory of the Subject Invention. The specific royalty rate and other terms of license shall be negotiated promptly in good faith and in conformance with the laws of the United States.

Article 7. Background Patent(s)

7.00 Laboratory Background Patent(s): Laboratory has filed patent application(s), or is the assignee of issued patent(s) which contain(s) claims that are related to research contemplated under this Agreement. No license(s) to this/these patent applications or issue patents is/are granted under this Agreement, and this/these application(s) and any continuations to it/them are specifically excluded from the definitions of "Subject Invention" contained in this Agreement.

7.01 Cooperator Background Patent(s): Cooperator has filed patent application(s), or is the assignee of issued patent(s) which contain(s) claims that are related to research contemplated under this Agreement. No license(s) to this/these patent applications or issue patents is/are granted under this Agreement, and this/these application(s) and any continuations to it/them are specifically excluded from the definitions of "Subject Invention" contained in this Agreement.

Article 8. Subject Data and Proprietary Information

8.00 Subject Data Ownership. Subject Data shall be jointly owned by the parties. Each party, upon request to the other party, shall have the right to review and to request delivery of all Subject Data, and delivery shall be made to the requesting party within two weeks of the request, except to the extent that such Subject Data are subject to a claim of confidentiality or privilege by a third party.

8.01 Proprietary Information/Confidential Information. Each party shall place a proprietary notice on all information it delivers to the other party under

this Agreement that it asserts is proprietary. The parties agree that any Proprietary Information or Confidential Information furnished by one party to the other party under this Agreement, or in contemplation of this Agreement, shall be used, reproduced and disclosed by the receiving party only for the purpose of carrying out this Agreement, and shall not be released by the receiving party to third parties unless consent to such release is obtained from the providing party.

8.02 Army limited-access database. Notwithstanding anything to the contrary in this Article, the existence of established CRADAs specifying areas of research and their total dollar amounts may be documented on limited access, password-protected websites of the U.S. Army Medical Research and Materiel Command (the parent organization of Laboratory), to provide the Command's leadership with a complete picture of military research efforts.

8.03 Laboratory Contractors. Cooperator acknowledges and agrees to allow Laboratory's disclosure of Cooperator's proprietary information to Laboratory's Contractors for the purposes of carrying out this Agreement. Laboratory agrees that it has or will ensure that its Contractors are under written obligation not to disclose Cooperator's proprietary information, except as required by law or court order, before Contractor employees have access to Cooperator's proprietary information under this Agreement.

8.04 Release Restrictions. Laboratory shall have the right to use all Subject Data for any Governmental purpose, but shall not release Subject Data publicly except: (i) Laboratory in reporting on the results of research may publish Subject Data in technical articles and other documents to the extent it determines to be appropriate; and (ii) Laboratory may release Subject Data where release is required by law or court order. The parties agree to confer prior to the publication of Subject Data to assure that no Proprietary Information is released and that patent rights are not jeopardized. Prior to submitting a manuscript for review which contains the results of the research under this Agreement, or prior to publication if no such review is made, each party shall be offered an ample opportunity to review any proposed manuscript and to file patent applications in a timely manner.

8.05 FDA Documents. If this Agreement involves a product regulated by the U.S. Food and Drug Administration (FDA), then the Cooperator or the U.S. Army Medical Research and Materiel Command, as appropriate, may file any required documentation with the FDA. In addition, the parties authorize and consent to allow each other or their contractors or agents access to, or to cross-reference, any documents filed with the FDA related to the product.

Article 9. Termination

9.00 Termination by Mutual Consent. Cooperator and Laboratory may elect to terminate this Agreement, or portions thereof, at any time by mutual consent.

9.01 Termination by Unilateral Action. Either party may unilaterally terminate this entire Agreement at any time by giving the other party written notice, not less than 30 days prior to the desired termination date.

9.02 Termination Procedures. In the event of termination, the parties shall specify the disposition of all property, patents and other results of work accomplished or in progress, arising from or performed under this Agreement by written notice. Upon receipt of a written termination notice, the parties shall not make any new commitments and shall, to the extent feasible, cancel all outstanding commitments that relate to this Agreement. Notwithstanding any other provision of this Agreement, any exclusive license entered into by the parties relating to this Agreement shall be simultaneously terminated unless the parties agree to retain such exclusive license.

Article 10. Disputes

10.00 Settlement. Any dispute arising under this Agreement which is not disposed of by agreement of the principal investigators shall be submitted jointly to the signatories of this Agreement. A joint decision of the signatories or their designees shall be the disposition of such dispute. However, nothing in this section shall prevent any party from pursuing any and all administrative and/or judicial remedies which may be allowable.

Article 11. Liability

11.00 Property. Neither party shall be responsible for damages to any property provided to, or acquired by, the other party pursuant to this Agreement.

11.01 No Warranty. The parties make no express or implied warranty as to any matter whatsoever, including the conditions of the research or any Invention or product, whether tangible or intangible, Made, or developed under this agreement, or the ownership, merchantability, or fitness for a particular purpose of the research or any Invention or product. The parties further make no warranty that the use of any invention or other intellectual property or product contributed, made or developed under this Agreement will not infringe any other United States or foreign patent or other intellectual property right. In no event will any party be liable to any other party for compensatory, punitive, exemplary or consequential damages.

Article 12. Miscellaneous

12.00 Governing Law. This Agreement shall be governed by the laws of the United States Government.

12.01 Export Control and Biological Select Agents and Toxins. The obligations of the parties to transfer technology to one or more other parties, provide technical information and reports to one or more other parties, and otherwise perform under this Agreement are contingent upon compliance with applicable United States export control laws and regulations. The transfer of certain technical data and commodities may require a license from a cognizant agency of the United States Government or written assurances by the Parties that the Parties shall not export technical data, computer software, or certain commodities to specified foreign countries without prior approval of an appropriate agency of the United States Government. The Parties do not, alone or collectively, represent that a license shall not be required, nor that, if required, it shall be issued. In addition, where applicable, the parties agree to fully comply with all laws, regulations, and guidelines governing biological select agents and toxins.

12.02 Independent Contractors. The relationship of the parties to this Agreement is that of independent contractors and not as agents of each other or as joint venturers or partners.

12.03 Use of Name or Endorsements. (a) The parties shall not use the name of the other party on any product or service which is directly or indirectly related to either this Agreement or any patent license or assignment agreement which implements this Agreement without the prior approval of the other party. (b) By entering into this Agreement, Laboratory does not directly or indirectly endorse any product or service provided, or to be provided, by Cooperator, its successors, assignees, or licensees. Cooperator shall not in any way imply that this Agreement is an endorsement of any such product or service. Press releases or other public releases of information shall be coordinated between the parties prior to release, except that the Laboratory may release the name of the Cooperator and the title of the research without prior approval from the Cooperator.

12.04 Survival of Specified Provisions. The rights specified in provisions of this Agreement covering Patent Rights, Subject Data and Proprietary Information, and Liability shall survive the termination or expiration of this Agreement.

12.05 Notices. All notices pertaining to or required by this Agreement shall be in writing and shall be signed by an authorized representative addressed as follows:

If to Cooperator: Georgetown University
Office of Technology Commercialization
3300 Whitehaven Street, N.W.
Harris Building, Suite 1500
Washington, DC 20007
Phone: 202-687-2702
Fax: 202-687-3111 (if by Fed Ex or courier)

Or use

Office of Technology Commercialization
Georgetown University
Box 571408
Washington, DC 20057-1408 (for US Mail)

If to Laboratory: USAMRIID
Business Plans and Programs Office
1425 Porter Street
Fort Detrick, MD 21702-5011
Phone: 301-619-6886 Fax: 301-619-8379

Any party may change such address by notice given to the other in the manner set forth above.

Article 13. Duration of Agreement and Effective Date

13.01 Effective Date. This Agreement shall enter into force as of the date it is signed by the last authorized representative of the parties.

13.02 Signature Execution. This Agreement may be executed in one or more counterparts by the parties by signature of a person having authority to bind the party, which may be by facsimile signature, each of which when executed and delivered, by facsimile transmission, mail, or email delivery, will be an original and all of which will constitute but one and the same Agreement.

13.03 Expiration Date. This Agreement will automatically expire two (2) years from effective date unless it is revised by written notice and mutual agreement.

IN WITNESS WHEREOF, the Parties have caused this agreement to be executed by their duly authorized representatives as follows:

For the Cooperator: Georgetown University
(Organization)

(Signature)

Claudia Cherney Stewart, Ph.D.

DATE 9/17/08 Vice President, Office of Technology Commercialization

For the U.S. Government: U. S. Army Medical Research Institute of Infectious Diseases

(Signature)

John P. Skvorak
Colonel, Veterinary Corps
Commanding

DATE 7 OCT 08

For the USAMRIID Principal Investigator:

I hereby acknowledge the terms and conditions of this Agreement:

DATE 24 Sept '08

(Signature)

Dr. Bradford Powell
(Printed Name)

(CRADA) APPENDIX A
STATEMENT OF WORK

Title: "Reanalysis and Functional Interpretation of Proteomics Data from Bacterial Cells under Simulated Growth Condition"

Background/Objectives:

Prior bacterial proteomics data needs to be reanalyzed due to the updates to the relevant bacterial protein databases and/or annotations, as well as accumulation of literature information regarding prior unknown genes. The objective of this collaboration is to use the integrated proteomics analysis system, iProXpress developed at PIR, coupled with the current TATRC- funded project, Pathogen Mining System, to facilitate the re-evaluation and functional interpretation and hypothesis formulation from the legacy proteomics data.

Prior 2DGE-MS proteomics data from Burkholderia strains grown under simulated host growth condition will be reanalyzed using iProXpress system for: 1) up-to-date functional assignment of bacterial protein annotations; 2) annotations of homologous proteins from other related pathogens of interests; 3) function and pathway analysis of the bacteria under given growth conditions.

Collaboration:

Laboratory agrees to:

- Provide MS data comprising protein lists and other information of relevance for matched data sets to be re-analyzed by the Pathogen Mining system.

Cooperator agrees to:

- Integrate all available annotations for proteins of Burkholderia and related bacteria into the iProXpress system, including biological pathways and experimental protein-protein interactions.
- Integrate into iProXpress the text mining results on pathogenesis proteins from the Pathogen Mining System.
- Incorporate the experimental Burkholderia proteomics data into the iProXpress system, and perform function and pathway analysis of the data.
- Enhance the iProXpress analysis interface based on the specific needs.

Document Classification for Mining Host Pathogen Protein-Protein Interactions

Guixian Xu[1,2,3,*], Lanlan Yin[1,*], Manabu Torii[4], Zhendong Niu[2], Cathy Wu[5], Zhangzhi Hu[5] and Hongfang Liu[1]

1. DBBB, Georgetown University Medical Center, Washington DC, USA
2. College of Computer Science, Beijing Institute of Technology, Beijing, China
3. College of Information Engineering, Central University for Nationalities, Beijing, China
4. ISIS Center, Georgetown University Medical Center, Washington DC, USA
5. PIR, Georgetown University Medical Center, Washington, DC, USA
{gx6,ly46,mt352,zh9,wuc,hl224}@georgetown.edu; zniu@bit.edu.cn

Abstract

Due to the heightened concern about bioterrorism and emerging/reemerging infectious diseases, a flood of molecular data about human pathogens has been generated and maintained in disparate databases. However, scientific findings regarding these pathogens and their host responses are buried in the growing volume of biomedical literature and there is an urgent need to mine information pertaining to pathogenesis-related proteins especially host-pathogen protein-protein interactions from literature. In this paper, we report our exploration of developing an automated system to identify MEDLINE abstracts referring to host-pathogen protein-protein interactions. An annotated corpus consisting of 1,360 MEDLINE abstracts was generated. With this corpus, we developed and evaluated document classification systems using support vector machines (SVMs). We also investigated the effects of feature selection using the information gain (IG) measure. Document classification systems were designed at two levels, abstract-level and sentence-level. We observed that feature selection was effective not only in reducing the dimensionality of features to build a compact system, but also in improving document classification performance. We also observed abstract-level systems and sentence-level systems yielded different classification of MEDLINE abstracts, and the combination of these systems could improve the overall document classification.

1. Introduction

Due to the heightened concern about bioterrorism and emerging/reemerging infectious diseases, there have been major initiatives for large-scale genomic and proteomic projects to study the basic biology and disease-causing mechanisms of human pathogens [1, 2]. As a result, a flood of molecular data is being generated, but important scientific discoveries regarding these pathogens and their host responses are often buried under the increasing volume of biomedical literature.

Over the years, biomedical literature mining advanced greatly. In this paper, our investigation focused on the development of an automated system to identify research articles describing pathogenicity and host-pathogen protein-protein interactions. Our goal is to facilitate literature-based curation of pathogenesis-related proteins in UniProt Knowledgebase (UniProtKB) [3] by incorporating pathogenesis information extracted from literature and promoting basic understanding of virulence and pathogenicity factors as well as host-interacting proteins of human pathogens. Such knowledge will facilitate the development of preventative and therapeutic strategies against human pathogens.

In the following, we first describe the research background and related work. The experimental method is introduced next. We then present the results and discussion, and conclude our work.

2. Background and related work

The task considered in this study is a special

* Equal contribution to the work.

case of identifying papers that describe protein-protein interactions (PPIs). There are several components in developing an automated literature mining system, including the construction of an annotated corpus, the selection of features and their representations, and the choice of machine learning algorithms. In the following, we present the research background and related work of each component.

2.1. Constructing annotated corpora from MEDLINE

One step towards constructing annotated corpora from MEDLINE is to select a subset of MEDLINE abstracts. There are different ways to obtain such subset. One approach is to use keyword search. For example, abstracts selected for the GENIA corpus were retrieved from MEDLINE using three MeSH terms, "human", "blood cell" and "transcription factor" [4]. An alternative way to obtain a subset is to exploit the use of existing biomedical databases. For example, in order to construct an annotated corpus for the Interaction Article Subtask at the second BioCreative workshop, contents of two existing interaction databases, namely IntAct and MINT, have been exploited [5]. After deriving such subset, domain experts can manually annotate them.

2.2. Feature representation/selection

In order to use machine learning methods, each document needs to be transformed into a feature representation, which is usually a feature vector. Commonly, features are based on words appearing in the document. Various feature selection techniques have been explored to overcome the high-dimensionality of word-based features [6, 7], e.g., Term Frequency (TF), TF * Inverse Document Frequency (IDF), Information Gain (IG), Mutual Information (MI), or chi-square statistics. In this paper, we explored IG for feature selection. IG represents the quantity of information in a feature with regard to class prediction on the base of presence/absence of the feature in a document. Let $\{c_i\}_{i=1}^{m}$ be a set of categories to be predicted. Then IG of feature *w* in a document collection is defined as follows:

$$G(w) = E - E_1 - E_2,$$

$$E = -\sum_{i=1}^{m} P(c_i)\log_2 P(c_i),$$

$$E_1 = -P(w)\sum_{i=1}^{m} P(c_i \mid w)\log_2 P(c_i \mid w),$$

$$E_2 = -P(\overline{w})\sum_{i=1}^{m} P(c_i \mid \overline{w})\log_2 P(c_i \mid \overline{w}),$$

where *E* is the entropy of the document collection; *m* represents the number of categories; $P(c) = \frac{N_c}{N}$ is occurrence probability of category c, where N represents the number of documents and N_c is the file numbers of class c; $P(w) = \frac{N_w}{N}$ and $P(\overline{w}) = \frac{N_{\overline{w}}}{N}$ are occurrence probabilities of presence and absence of *w*, N_w and $N_{\overline{w}}$ are the file numbers of including and not including feature w in the document collection; and finally $P(c \mid w) = \frac{N_{wc}}{N_w}$ and $P(c \mid \overline{w}) = \frac{N_{\overline{w}c}}{N_{\overline{w}}}$ are occurrence conditional probability of the category c on occurrence and absence of term *w*, where N_{wc} and $N_{\overline{w}c}$ are the file numbers of including and not including term w in class *c*[8]. It is assumed that the larger the IG value of a term is, the more important the term is in classifying documents.

2.3. Machine learning algorithms

A growing number of statistical and probabilistic machine learning algorithms have been applied to document classification, including K nearest neighbor, Bayesian approaches, decision trees, symbolic rule learning, and neural networks [9-12]. Here, we chose Support Vector Machines (SVMs), a supervised learning algorithm proposed by Vladimir Vapnik and his co-workers [13, 14]. It has been widely used for text mining and achieved promising results. Given a training set with n class-labeled instances, (x_1, y_1), (x_2, y_2), ..., (x_i, y_i), …, (x_n, y_n), where x_i is a feature vector for the i-th instance and $y_i \in \{+1, -1\}$ indicates the class, an SVM classifier learns a hyper-plane as a decision boundary in the feature space. The class of an unlabelled instance x is determined by on which side of the hyperplane x lies. The purpose of training SVM classifiers is to find a hyperplane that has the maximum margin to separate the two classes [16-18].

3. Method

Figure 1 illustrates the overall data flow of the classification system. It consists of several steps including i) generating annotated MEDLINE abstracts, where each abstract was annotated either positive or negative (e.g., +1 or -1) based on its relevance to host-pathogen protein-protein interactions (PH-PPI), ii) conducting machine learning experiments to evaluate different kinds of feature representations and feature selection methods, and iii) implementing a system that assigns confidence scores to abstracts based on their PH-PPI relevance.

3.1. Generation of an annotated corpus

The annotated corpus was generated from two different sources. One was from UniProtKB database where the PH-PPI information is annotated for the protein entries and the relevant MEDLINE abstracts are cited. If a cited abstract contains an interaction pair consisting of one host protein and one pathogen protein, it is considered as positive. The other source was from PubMed, from which a set of MEDLINE abstracts was retrieved using keyword searches. Two domain experts reviewed and manually annotated this set, and categorized the abstracts as positive or negative. Additionally, for positive abstracts sentences describing the interactions were highlighted.

3.2. Machine learning

Instead of classifying a document as PH-PPI relevant or not, the machine learning task considered here is to rank a set of documents according to their PH-PPI relevance. We defined two machine learning tasks. One task is at abstract level (ALT), which uses the abstracts to build a system to rank a set of abstracts according to their PH-PPI relevance. The other is on sentence level (SLT) which ranks all sentences in abstracts by considering titles and highlighted sentences in positive abstracts as positive and all sentences in negative abstracts as negative. The ranking of a set of abstracts can then be obtained according to the rank of the most relevant sentence in an abstract.

3.2.1. Feature representation/selection

We normalized the text by changing nouns in plural forms into singular forms, verbs in past tense into present tense, and replacing nouns and adjectives by their corresponding verbs based on the SPECIALIST lexicon, a component in the Unified Medical Language System (UMLS). We also replaced punctuation marks with spaces and changed uppercase letters to lowercase letters.

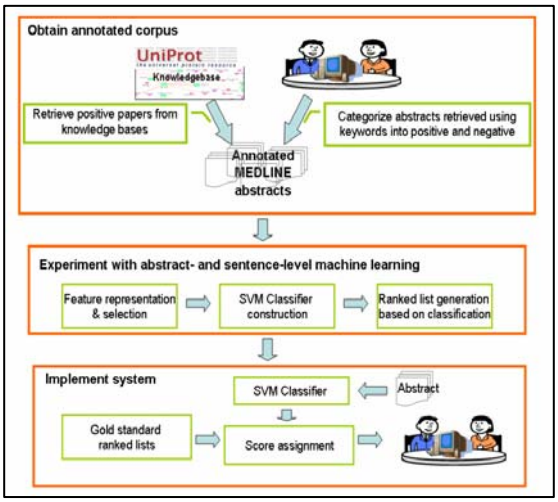


Figure 1. Overall architecture of the study.

After normalization, we used unigrams and bigrams as features, and the frequencies of unigrams and bigrams as their corresponding feature values. To reduce the dimensionality of the feature space, we used information gain to select features with high IG values. Note that we did not remove features that are stop or rare words in this work.

3.2.2. Machine learning algorithms

We used the SVM light package and chose a linear function as the kernel [13]. We also experimented with other types of kernels such as polynomial or radial basis function (RBF), but observed no performance improvement.

3.2.3. Experiments

The experiments were designed to i) compare IG feature selection (IG-FS) with no feature selection (NO-FS), and ii) compare ALT and SLT. We used 100 runs of 10-fold cross validation. For each run, the same 10 fold partitions were used for the following four settings: (IG-FS, ALT), (IG-FS, SLT), (NO-FS, ALT), and (NO-FS, SLT). For each setting, we obtained a ranked list consisting of abstracts in the annotated corpus ranked according to the results of the 10-fold cross validation experiment. The performance was then measured using true positive rate (TPR): given rank threshold P and ranked list L, TPR(P, L) is defined as the ratio of the number of true positives ranked as top P in L to P. We selected 18 different rank thresholds: from 10 to 90 (incremented by 10) and from 100 to 500 (incremented by 50). In case of IG-FS, we set 20 IG thresholds: 0 to 0.0009

(incremented by 0.0001) and from 0.001 to 0.01 (incremented by 0.001). For each IG threshold, we ignored all features with IG values less than the threshold when constructing the systems. The average TPR of 100 runs for each setting was computed to compare the performance. Confidence intervals at 95% Confidence Level were also computed [15].

3.3. System implementation

As we have discussed, the machine learning task considered here is to rank a set of documents according to their PH-PPI relevance. In order to judge the PH-PPI relevance for any given abstract, we used the following method:

i) obtain N score lists by executing N runs of 10-fold cross validation using the corpus as described in Section 3.2.3 where scores were ones assigned by SVM classifiers,

ii) build a SVM classifier C with all instances in the corpus,

iii) for a new abstract, use classifier C to obtain score S,

iv) for each score list that was obtained in i) compute the percentage of instances that are positive among the instances with scores larger than S, and

v) average the above percentage over N score lists and display the percentage as the relevance score. The higher the score, the more relevant the abstract.

To test the effectiveness of the proposed method, we used one run of 10-fold cross validation and measured TPRs for a given relevance score threshold.

4. Result and discussion

Most pathogen protein-protein interaction (PPI) information annotated in knowledgebases is for viral proteins or PPI within bacteria. We obtained less than 50 positive abstracts on specific bacterial pathogen-human host PPI from knowledge bases such as UniProtKB/Swiss-Prot and, IntAct, Brucella Bioinformatics Portal (BBP). Using key words "bacterial", "host", "pathogen", and "interaction", we retrieved around 214,000 abstracts, and we obtained 1,225 negative abstracts and 99 positive abstracts after manual annotation. Merging the two sets, the annotated corpus consists of 1,225 negative abstracts and 135 positive ones.

Figure 2 shows the relationship between IG threshold and TPR averaged over 100 runs. The IG threshold of 0 corresponds to no feature selection (NO-FS). From Figure 2, we can see that for IG

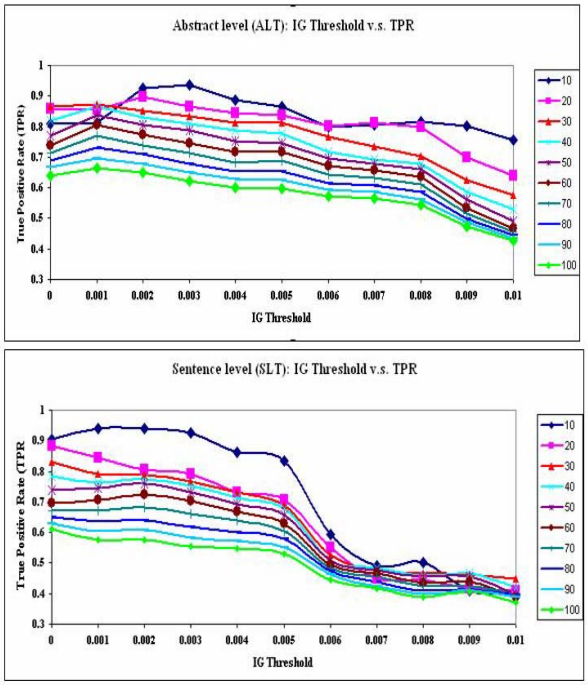


Figure 2. The relationship between IG threshold and TPR averaged over 100 runs in (IG-TF, ALT) and (IG-FS, SLT).

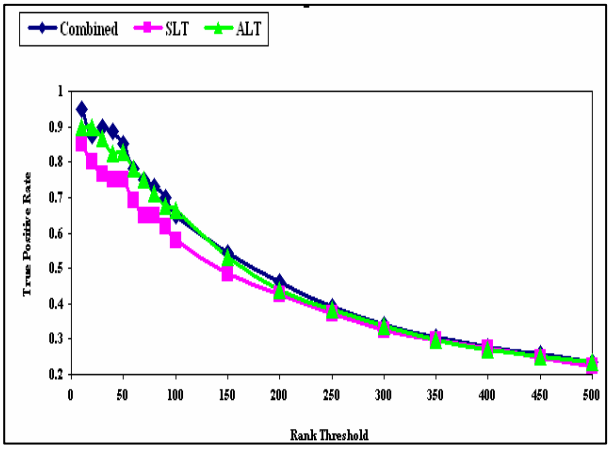


Figure 3. Combination result of (IG-FS, ALT)-0.001 and (IG-FS, SLT)-0.001.

thresholds between 0.001 and 0.005, the TPRs are comparable to the one without feature selection (i.e. NO-FS). However, the number of features used for classifiers with feature selection decreases dramatically. For example, in (IG-FS, ALT) with threshold 0.002 and (IG-FS, SLT) with threshold 0.001, the number of features after feature selection is reduced to only 10% (around 10,000) of the original (over 100,000).

Table 1. The detailed TPRs with the corresponding 95% confidence intervals computed from 100 runs for (IG-FS, ALT), (IG-FS, SLT), (NO-FS, ALT) with IG threshold 0.002, and (NO-FS, SLT) with IG threshold 0.001. RT stands for rank threshold.

RT	NO-FS		IG-FS	
	ALT	SLT	ALT(0.002)	SLT (0.001)
10	0.81 (0.794, 0.827)	0.905 (0.899, 0.911)	0.926 (0.915,0.937)	0.941 (0.930,0.952)
20	0.857 (0.852, 0.862)	0.883 (0.875, 0.891)	0.898 (0.890,0.906)	0.844 (0.834,0.854)
30	0.867 (0.862, 0.873)	0.832 (0.825, 0.839)	0.852 (0.845,0.859)	0.79 (0.782, 0.798)
40	0.819 (0.812, 0.826)	0.786 (0.779, 0.793)	0.83 (0.823,0.837)	0.764 (0.757, 0.771)
50	0.768 (0.762, 0.774)	0.737 (0.731, 0.744)	0.807 (0.801,0.813)	0.745 (0.739, 0.751)
60	0.74 (0.735, 0.745)	0.697 (0.692, 0.702)	0.775 (0.770, 0.780)	0.706 (0.700, 0.712)
70	0.715 (0.710, 0.720)	0.67 (0.665, 0.675)	0.738 (0.733, 0.743)	0.67 (0.665, 0.675)
80	0.69 (0.686, 0.694)	0.65 (0.646, 0.654)	0.71 (0.705, 0.715)	0.637 (0.633, 0.642)
90	0.666 (0.662, 0.67)	0.629 (0.625, 0.633)	0.679 (0.674, 0.684)	0.604 (0.600, 0.608)
100	0.639 (0.635, 0.643)	0.611 (0.6068,0.6152)	0.649 (0.645, 0.653)	0.577 (0.574, 0.581)
150	0.515 (0.513, 0.517)	0.514 (0.511, 0.517)	0.522 (0.519, 0.525)	0.491 (0.488,0.494)
200	0.431 (0.429, 0.433)	0.431 (0.429, 0.433)	0.438 (0.436, 0.440)	0.429 (0.427, 0.431)
250	0.377 (0.376, 0.379)	0.371 (0.369, 0.373)	0.378 (0.376, 0.380)	0.379 (0.377, 0.381)
300	0.336 (0.335, 0.337)	0.33 (0.329, 0.331)	0.334 (0.332, 0.336)	0.336 (0.334, 0.338)
350	0.303 (0.302, 0.304)	0.301 (0.300, 0.302)	0.3 (0.299, 0.301)	0.301 (0.300, 0.302)
400	0.278 (0.277, 0.279)	0.276 (0.275, 0.277)	0.273 (0.272, 0.274)	0.272 (0.271, 0.273)
450	0.257 (0.256, 0.258)	0.255 (0.254, 0.256)	0.251 (0.250, 0.252)	0.249 (0.248, 0.250)
500	0.238 (0.237, 0.239)	0.236 (0.235, 0.237)	0.232 (0.231, 0.233)	0.229 (0.228, 0.230)

Table 1 shows the detailed results of four settings: (NO-FS, ALT), (NO-FS, SLT), (IG-FS, ALT) with IG threshold 0.002, and (IG-FS, SLT) with IG threshold 0.001. For example, among top 50 abstracts, there are 76.8%, 73.7%, 80.7%, and 74.5% of the abstracts are positive for (NO-FS, ALT), (NO-FS, SLT), (IG-FS, ALT), and (IG-FS, LT), respectively. The average TPRs usually decrease when the rank thresholds increase. The performance of sentence-level systems is comparable to that of abstract-level systems when the rank threshold is small (e.g., 10 or 20). When the rank threshold (e.g., > 20) is large, abstract-level systems tend to perform better.

Table 2 shows the performance of the true positive rate when implementing the system using (IG-FS, ALT) with IG threshold 0.002 and the number of runs as 5. Given a relevance score threshold 0.5, the true positive rate is 50.7% which indicates that if an abstract receives a relevance score of larger than 0.5, the chance of the abstract to be positive is 50.7%.

Table 2. The performance of the implementation.

Threshold	Total	Positive	TPR
0	1,360	135	0.099
0.1	1,185	118	0.099
0.2	519	106	0.204
0.3	304	93	0.306
0.4	207	82	0.396
0.5	136	69	0.507
0.6	96	63	0.656
0.7	69	52	0.754
0.8	41	30	0.732
0.9	8	7	0.875

Even sentence-level systems perform inferior to abstract-level systems, but one advantage of them is that sentences describing protein interactions are automatically highlighted. We can highlight sentences

(and titles) yielding the highest ranks among sentences within the abstract when presenting the results to end-users. For example, for (IG-FS, SLT) with IG threshold 0.001, the average number of positive abstracts is 17 (or 37) among the top 20 (or 50) abstracts. Among those positive abstracts, an average of 13 (or 26) abstracts have the highlighted sentences ranked as the highest among all sentences in the corresponding abstract by the sentence-level systems, and an average of 16 (or 33) abstracts have the highlighted sentences ranked as the highest or the second highest.

We also noticed that sentence-level systems and abstract-level systems behave differently. The finding is consistent with the work of Ding et al where different text units (e.g., abstracts, sentences, or phrases) were investigated for information retrieval [16]. Given rank threshold 10, and IG threshold 0.001, the average number of overlapped true positives between sentence-level and abstract-level systems is around 4. We checked the combination of sentence-level and abstract-level systems by averaging the ranks of sentence-level and abstract-level. Figure 3 shows the result. There is some improvement of the performance after combination.

5. Conclusion

We have reported a study of constructing an automated system that can detect the host pathogen protein-protein interaction relevance of MEDLINE abstracts. The results indicated that feature selection can reduce the number of features at least 10 folds with no or little sacrifice of performance. Additionally, the majority of the highlighted sentences are ranked as the first or second among all sentences in the corresponding abstracts. We conclude that automated systems can be built for retrieving abstracts and highlighting sentences based on their relevance to host pathogen protein-protein interaction.

6. Acknowledgements

This work was supported by US Army TATRC #W81XWH0720112 and NSF IIS-0639092.

References

[1] CG Zhang, BA Chromy and SL McCutchen-Maloney. Host–pathogen interactions: a proteomic view. *Expert Review of Proteomics*, 2(2):187-202, 2005.

[2] K Nomura, S DebRoy, YH Lee, N Pumplin, J Jones and SY He. A Bacterial Virulence Protein Suppresses Host Innate Immunity to Cause Plant Disease. *Science*, 313 (5784): 220-223, 2006.

[3] CH Wu, R Apweiler, A Bairoch, DA Natale, WC Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, M J Martin, R Mazumder, C O'Donovan, N Redaschi and Baris Suzek The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research* 2006.

[4] J.-D. Kim, T. Ohta, Y. Tateisi and J. Tsujii. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*. 19(Suppl. 1): i180-i182, 2003.

[5] M Krallinger and A Valencia. Evaluating the Detection and Ranking of Protein Interaction Relevant Articles: the BioCreative Challenge Interaction Article Sub-task (IAS). In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 29-39, 2007.

[6] Y Yang and J Pederson. A Comparative Study on Feature Selection in Text Categorization". *Proceedings of the fourteenth International Conference on Machine Learning*, Pages 412-420, 1997.

[7] E Cantú-Paz, S Newsam and C Kamath. Feature selection in scientific applications. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Pages 788 – 793, 2004.

[8] K Machová and A Szaboová. Statistical Methods in Key Words Generation from Text Documents. *In 5th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence and Informatics*, pages 435-446, 2007.

[9] Y Yang and CG Chute. An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems (TOIS)*, vol 12, Pages 252 – 277, 1994.

[10] DD Lewis, M Ringuette. A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 81-93, 1994.

[11] WW Cohen and Y Singer. Context-Sensitive Learning Methods for Text Categorization. *ACM Transactions on Information Systems (TOIS)*, Vol 17, Pages 141 – 173, 1999.

[12] ED Wiener, JO Pedersen and AS Weigend. A neural network approach to topic spotting. In *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, Pages 317-332, 1995.

[13] VN Vapnik. Statistical Learning theory[M].1998.

[14] HF Liu, C Wu. A Study of Text Categorization for Model Organism Databases. *HLT-NAACL 2004 Workshop: Biolink 2004, Linking Biological Literature, Ontologies and Databases*, pages 25-32, 2004.

[15] U Hahn, M Romacker and S Schulz. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. *Pac Symp Biocomput*, pages 338-349, 2002.

[16] J Ding, D Berleant, D Nettleton, and E Wurtele. Mining MEDLINE: abstracts, sentences, or phrases? *Pac Symp Biocomput*, pages 326-337, 2002.

iProLINK: A Framework for Linking Text Mining with Ontology and Systems Biology

Zhang-Zhi Hu[1*], K. Bretonnel Cohen[2], Lynette Hirschman[2], Alfonso Valencia[3], Hongfang Liu[4], Michelle G. Giglio[5], Cathy H. Wu[1]

[1]Protein Information Resource, [4]Department of Biostatistics, Biomathematics and Bioinformatics, Georgetown University Medical Center, Washington DC; [2]Information Technology Center, The MITRE Corporation; [3]Spanish National Cancer Research Centre (CNIO), Spain; [5]University of Maryland School of Medicine, Baltimore, MD

{zh9, hl224, wuc}@georgetown.edu; {kbcohen, lynette}@mitre.org; valencia@cnio.es; mgiglio@som.umaryland.edu

Abstract

The ever-increasing scientific literature and the exponential growth of large-scale molecular data have prompted active research in biological text mining to facilitate literature-based curation of molecular databases. Meanwhile, systems biology and bio-ontologies are emerging as critical tools in biological research where complex data in disparate resources are generated, integrated and analyzed. Both rely on literature for data annotation and analysis. The challenges facing us are to develop broadly utilized text mining tools and systems, and to bring together developer and user communities for system development and evaluation. We describe a framework for linking text mining tools with ontology and systems biology, extending from a previously developed text mining resource, iProLINK. We focus on molecular and ontological resources, including genes/proteins, protein-protein interaction (PPI), and Protein Ontology. The framework consists of two major components: a user interface for text mining of PPI from an integrated tool server and software modules to allow text mining outputs to be created, ranked, and used by the community. Use cases are presented for assessing the gaps and making recommendations for future development.

1. Introduction: current status of text mining as an enabling tool for biology

The biological literature represents the repository of biological knowledge. As biology becomes more dependent on information technology, there has been an explosion of computable resources and databases [1], e.g. GenBank, UniProt, model organism databases, and systems biology databases, e.g., Reactome, KEGG, that capture much of the structured information on sequence and functional data. It becomes critical to link these data sources to their associated context, e.g., experimental methods and evidence. Such information is largely buried in the literature and it has become prohibitively expensive for curators to keep up with its growth.

1.1. Text mining resource development

Most of the work in biomedical text mining over the past decade has focused on solving specific problems, often using task-tailored and private datasets, which were rarely reused. As more research groups began to make resources publicly available, there have been a number of projects, initiatives and organizations dedicated to building and providing access to biomedical text mining resources, such as those listed at the National Center for Text Mining at UK (http://www.nactem.ac.uk) and Text Mining Group at the Center for Computational Pharmacology (http://compbio.uchsc.edu/ccp/corpora).

Researchers at PIR have contributed to this effort by developing a literature mining resource, iProLINK, to support text mining and NLP research for bibliography mapping (references cited in a protein entry), annotation extraction, entity recognition and protein ontology development [2]. The data sources for bibliography mapping and feature evidence attribution include mapped citations and annotation-tagged literature corpora [3]. The data sources for entity recognition and ontology development include protein name dictionaries and protein name-tagged literature corpora along with tagging guidelines [4]. These curated corpora have been used for training and benchmarking text mining tools such as RLIMS-P, an information extraction tool for protein phosphorylation [5]. iProLINK also provides the online BioThesaurus, a large collection of gene/protein names with UniProt entry associations [6].

1.2. Text mining critical evaluations

* Corresponding author.

As the BioCreative [7, 8] and TREC Genomics track [9] evaluations have shown, common evaluations are important to create an active research community and to accelerate the research progress. There have been two BioCreative workshops to date, with 27 groups participating in the first [7], and 44 groups participating in the second [8]. These workshops have focused on tasks relevant to the biological curation community, including identification of gene mention (GM) and gene normalization (GN), and on more advanced tasks. For BioCreative I, the focus was on functional annotation, including linkage of evidence passages to support GO annotations for proteins in full text articles. For BioCreative II, the advanced task focused on extraction of protein-protein interaction (PPI) information, using "gold standard" data provided by the MINT and IntAct databases. The BioCreative evaluations have driven progress in biomedical text mining and have led to release of annotated data collections for further evaluation (http://BioCreative.sourceforge.net).

1.3. Text mining tool integration

It has been observed that "accurate and diverse" tools targeting the same application area can make a combination system outperform a single constituent tool [10, 11]. For example, Si et al. [12] combined systems that participated in the JNLPBA shared task (recognition of five types of entities in abstracts), and reported excellent performance using Conditional Random Fields (CRFs). Similarly [13, 14] reported results obtained by combining 21 systems from the BioCreative II GM task, and reported an F-measure over 90% using CRFs.

A major accomplishment of BioCreative II was the establishment of BioCreative MetaServer (BCMS, http://bcms.bioinfo.cnio.es/) [15], a prototype platform that combines text mining services from multiple groups, currently covers some major tasks from BioCreative II, including GM/GN, taxon classification and PPI identification, and provides annotations from 13 servers for the BioCreative corpus of MEDLINE abstracts.

1.4. Text mining standards development

Common standards for data exchange and tool integration are critical for text mining. Currently there is a lack of formal standards and candidates for de facto standards are not widely accepted at this time. The first concrete proposal for a data exchange standard for biomedical text processing was GPML, the GENIA Project Mark-up Language [16]. A corpus annotated in this format has been released in multiple revisions and has experienced significant acceptance in the text mining community [17], but tool producers have not embraced it as an output format. For the tool integration, there has been considerable amount of interest in the Unstructured Information Management Architecture (UIMA) [18-21], but it is not considered the de facto standard for tool integration yet. A meeting held in conjunction with the recent BioNLP 2008 workshop concluded that there was little hope for convergence on a common format in the near future, and that the best that could be hoped for at this time with respect to corpora and data exchange is that corpus builders produce formats that can be interconverted—no small feat in itself [22].

1.5. Motivation for a community framework

Even with advancements in tool and system development and the growing collaborative efforts of the text mining community, literature mining tools are still not broadly used by biological communities. Such a gap is partly due to intrinsic complexity of biological text for mining, and partly to the lack of close interactions between the text mining and the user communities, represented by biology researchers and curators.

BioCreative I and II focused on critical assessment of text mining tool performance on individual tasks involved in the overall molecular data curation process. The next step is to link these tools together to provide an environment that supports end users. The communities represented by biologists/curators and tool developers can be brought together by a common interface and through community workshops. In this paper, we describe an extended iProLINK framework that aims to link the three communities, allowing text mining tools to be evaluated and adopted by the broad communities. This work builds on four threads of research: the previous iProLINK text mining resource; BioCreative evaluations; tools and data resources developed under BioCreative, in particular the vision of a MetaServer to provide text mining services to users; and work at PIR focused on building a framework for the capture of PPI, including post-translational modifications (PTM). We present several case studies that illustrate the mutual benefit each community can gain from the others.

2. Linking text mining with ontology and systems biology: a basic framework

2.1. iProLINK framework

An overview of the iProLINK framework is shown in Figure 1. It contains two major components: text mining tools, and the interface that links the text mining to ontology and systems biology communities. Text mining tools are integrated into a metaserver that will generate text mining results, and the user interface will display ranked outputs (circle #1) and the visualized protein networks (#2) based on the output. The interface also allows users/curators to curate the text mining results (#1) and make assertions on the extracted knowledge. The curated information is used for or captured in ontologies (e.g. Protein Ontology) (#3) and

knowledgebases (#4), and is also saved in a curated literature corpus (#6) used for improving the text mining output ranking (#7) and for enhancing text mining tool development (#8). The systems biology data can also be used to help assertion of the text mining results (#5).

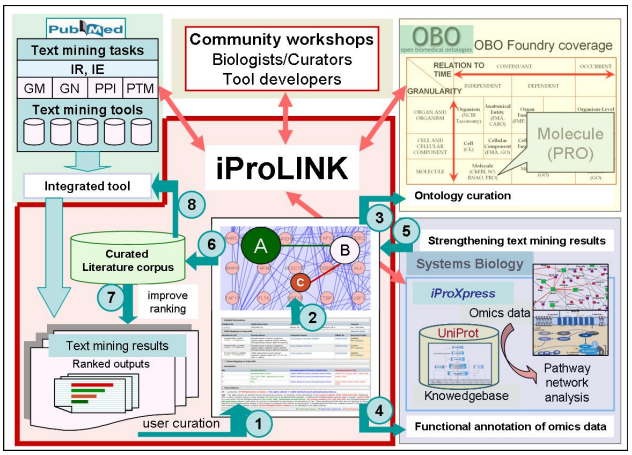


Figure 1. Overview of the iProLINK framework

2.2. Linking text mining, ontology and systems biology for protein-protein interactions

PPI generally refers to physical associations of two protein objects, stable or transient, such as in protein complexes or in signaling cascades. There are many types of PPIs; in this context, we define PPI as protein pairs with either direct or indirect associations such as through intermediate steps.

Text mining. The text mining tasks for iProLINK include integration of tools, presently covering gene or protein mention, gene or protein normalization, and information retrieval and extraction of PPI, including PTMs such as phosphorylation (an interaction between a protein substrate and a protein kinase). There are a number of tools for these tasks, including those participating in the BioCreative I and II challenge evaluations, and others such as RLIMS-P.

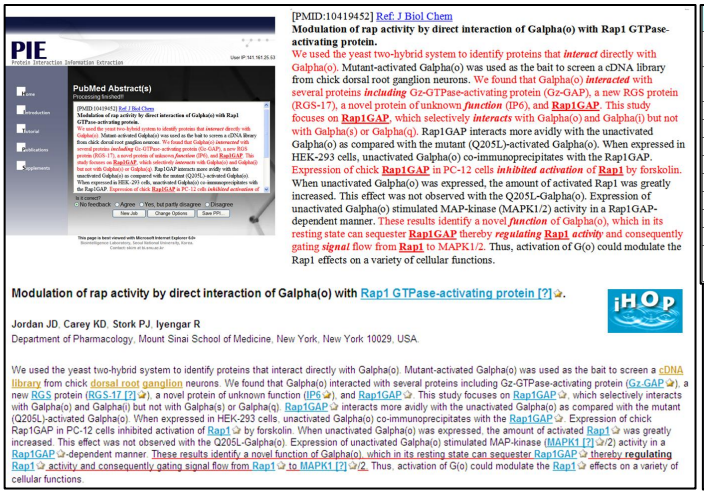
Ontology. Open Biological Ontologies (OBO) Foundry is a collaborative effort for coordinating various biological ontology development projects and for fostering common standards in OBO development [23]. The curation of the content of ontologies, especially those related to genes or proteins, e.g. specific splice or modified forms of gene products in Protein Ontology (PRO) [24], relies heavily on literature information. In particular, protein PTM and PPI text mining will help annotate protein nodes (terms) by identifying specific phosphorylated forms and adding PPI information as attributes to PRO forms.

Systems biology. Molecular databases represent structured knowledge of genes/proteins, such as UniProt, and biological pathway and PPI databases. Annotation of those databases and utilization of the annotations for large-scale omics data analysis are an integral part of systems biology, e.g., iProXpress, an expression analysis system for systems biology [25]. Text mining results can be used to infer or add more evidence to pathway and network analysis results derived from systems biology data; conversely, large-scale data can be used to support the text mining results of PPI information.

3. iProLINK use case analysis

3.1. PPI text mining for generation of protein networks

There are several PPI text mining tools, such as PIE [26] and iHOP [27], both as part of the BCMS. We use these two tools to illustrate PPI text mining results and how they can be used for generation of protein networks. As shown in Figure 2, the tools typically highlight or underline sentences containing the PPI, with protein pairs and words for relations highlighted (bold or colors). There are 11 pairs of PPI instances in this abstract, including the title. Most (8/11) are detected by one or the other tool, and most (9/11) are direct PPIs.



Protein 1	Protein 2	Tools
Galpha(o)	Rap1 GTPase-activating protein	n/a
Galpha(o)	Gz-GTPase-activating protein (Gz-GAP)	PIE
Galpha(o)	RGS protein (RGS-17)	PIE
Galpha(o)	IP6	PIE
Galpha(o)	Rap1GAP	PIE
Rap1GAP	Galpha(o)	PIE
Rap1GAP	Galpha(i)	PIE
Galpha(o)	MAP-kinase (MAPK1/2)	n/a
Galpha(o)	Rap1	n/a
Rap1	MAP-kinase (MAPK1/2)	iHOP
Rap1GAP	Rap1	PIE iHOP

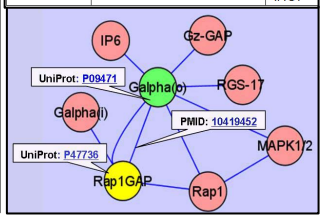


Figure 2. PPI text mining results for the construction of protein network

The visualized PPI network allows users to more efficiently mine proteins of interest and their interacting partners. Based on the binary relations (edge) between interacting partners (node), we used Cytoscape [28] to display these mined PPIs in a single protein network (Figure 2, lower right). It shows that Galpha(o) is a

major hub protein that interacts with six other proteins directly or indirectly. Rap1GAP is another important protein that interacts with three other proteins. The UniProt IDs for the protein nodes are displayed with mouse-over, and the text evidence for relations (edges) between protein nodes is also visualized by mouse-over (PMID in this case). The protein networks can also be built from multiple abstracts either through batch retrieval (section 3.3) or by gene/protein name searches. The latter would be a more useful feature in analyzing PPI of particular proteins based on PubMed searches.

The essential requirements for the interface in PPI text mining and protein network generation are to 1) provide ranking of PPI outputs based on scores or confidence levels for each protein pairs; 2) support user/curator feedback on the output ranking and content, and an ability to save the output in standard data formats compatible to other software tools such as Cytoscape and OBO editor; and 3) display the protein nodes and edges with weightings and evidence attributions.

3.2. PTM text mining for Protein Ontology form curation

The Protein Ontology is designed to describe the relationships of proteins and protein evolutionary classes, to delineate the multiple protein forms of a gene locus, and to interconnect existing ontologies [24]. Multiple protein forms include splice isoforms and various PTMs. Knowledge of protein splice forms and modifications are mostly embedded in the literature, thus text mining of such information greatly facilitates the curation of PRO nodes (terms) and relations. Protein phosphorylation is a common type of PTM, and proteins with distinct phosphorylated residue(s) represent unique protein forms. RLIMS-P is designed to extract the three protein phosphorylation objects: kinase, substrate and the phosphorylation sites/residues. The kinase and substrate interaction is a special case of PPI that can be mined by text mining tools, such as PIE. However, RLIMS-P also extracts phosphorylation sites, useful for PRO curation.

Figure 3 shows the output of the RLIMS-P extracted PPI and phosphorylation sites (PMID: 18003885), which can be directly used for curation of the protein node, RUNX1, a transcription factor. RLIMS-P outputs contain a summary table for the extracted PPI and evidence-tagged sentences in the abstract. One of the 11 isoforms, AML-1G, of human RUNX1 is described in PRO format as being phosphorylated at Ser 48, 303, and 424; the specific PTM type (phosphorylation at L-serine) is annotated using the PSI-MOD ontology (MOD:00046) (Figure 3). Experimental PPI information can also be used for annotating properties to protein forms in PRO, e.g., the associated functions of the phosphorylated form of RUNX1 in this paper can be annotated for AML-1G, e.g., "increases transactivation potency and stimulates cell proliferation". The RLIMS-P outputs need to be saved in standard formats, such as OWL or OBO, for protein network display and PRO curation.

Protein 1	Protein 2	Rel.	PMID	Tool
Galpha(o)	Rap1GAP	I	10419452	PIE
Cdc2	Rap1GAP	P	8048970	RLIMSP
Cdc2	BAD	P	12049737	RLIMSP PIE
BAD	14-3-3	I	12049737	RLIMSP PIE
			10949026	RLIMSP PIE
BAD	Bcl-xL	I	8929531	PIE

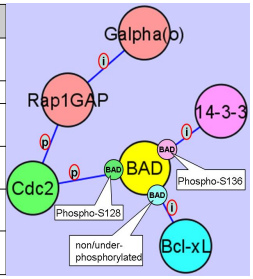


Figure 4. Text mining summary and network generation of PPI, including general "Interaction" (I) and protein phosphorlyation (P)

3.3. PPI text mining for systems biology

Systems biology data include gene/protein databases and large-scale omics data repositories. Annotation and analysis of systems biology data can benefit from PPI text mining. The protein network in Figure 2 contains the Rap1-MAPK pathway, which is modulated by Gα(o)-Rap1GAP interaction. Other papers describe the activation of Rap1GAP through phosphorylation by Cdc2 (CDK1), which also phosphorylates the BAD protein at distinct site (Ser128) (Figure 4). Interestingly, distinct forms of BAD interact with different partners.

When combining PPI mining results from Figure 2 and 4, a larger protein network can be generated, showing four highly-connected protein nodes—Gα(o), Rap1GAP, Cdc2 and BAD (Figure 5A). Compared to a pathway diagram

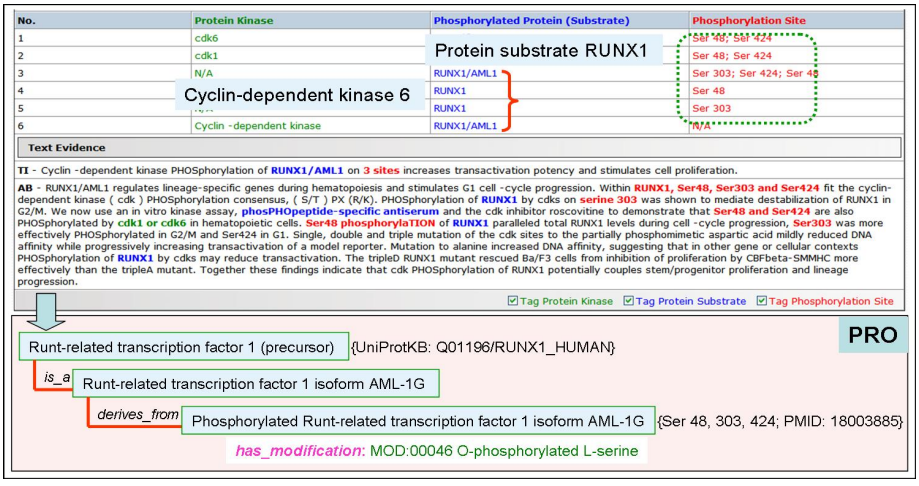


Figure 3. RLIMS-P text mining for Protein Ontology curation

based on the analysis of a proteomics dataset [29] (Figure 5B), this text mining-based PPI network graph not only provides literature evidence for the interactions shown in the pathway map (e.g., GNAO2-Rap1GAP, Rap1GAP-Rap1), but also reveals a missing interacting protein pair (Cdc2-Rap1GAP) in the pathway (red dashed arrow), as well as missing partners of BAD protein (14-3-3 and Bcl-xL).

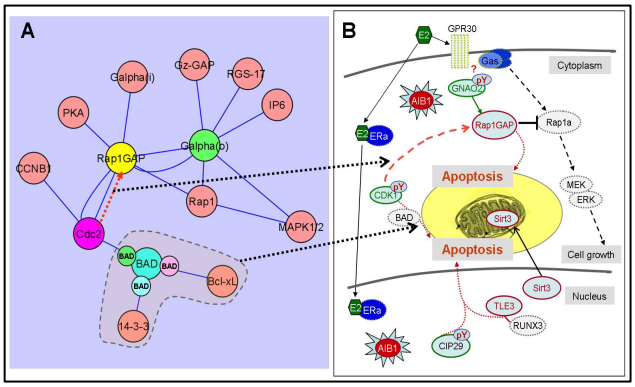


Figure 5. Text mining of PPI (A) for annotation and analysis of systems biology data (B)

3.4. PPI text mining supported by systems biology data

Systems biology data can also strengthen PPI text mining results. Figure 6 shows an example where PPI proteomic data from large-scale immunoprecipitation are linked to text mining results. The Sp1-p38 interaction from a proteomics experiment was deposited in IntAct, one of the PPI and pathway databases integrated into the iProXpress underlying data warehouse. This information supports the protein network derived from text mining, showing p38-Sp1 interaction and activation of filamin-A.

The display of protein networks will allow linkage of protein nodes to pathway maps or high-throughput PPI data from molecular databases. Alternatively, saved text mining outputs can also be integrated into users' pathway and network analysis pipeline.

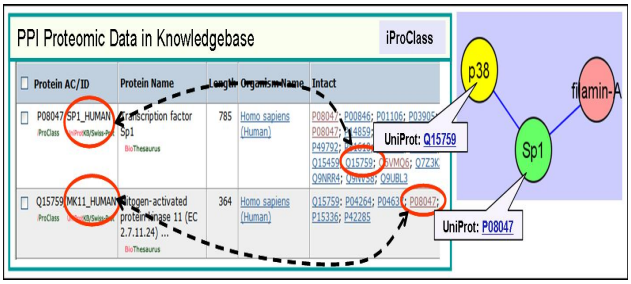


Figure 6. Systems biology data support the text mining of Sp1-p38 interaction (PMID: 12324467)

4. Future work

From above case studies, we have identified major gaps between the text mining and the ontology and systems biology communities that need to be addressed:

Standards development. Text mining standards include those of data exchange and tool integration. Tool integration involves issues such as process control and preserving state information as well as a mechanism for exchanging data. Standards must also support data exchange, including both syntactic standards (e.g., XML or SGML tags) and semantic standards – perhaps based on widely accepted biological resources, such as EntrezGene and UniProt.

User interface requirements. The web interface is a major component of the iProLINK framework for the communities. The new interface will allow biologists to browse, curate, and save the text mined PPI/PTM information. The interface should provide several key functionalities: 1) The output from multiple text mining tools should be ranked, and the display of protein network and associated text evidence should be weighted; 2) Users should be able to edit the text mining results, and the asserted knowledge should be saved in standard or convertible formats for use by different communities; 3) The interface should be simple to users with customizable options and views.

Usability testing. A major activity of iProLINK will be to facilitate interactions between text mining and user communities through annual workshops including joint workshops with existing activities, such as BioCreative and International BioCuration Meetings. An annotation workshop will allow database curators to experiment with integrating multiple text mining tools into their workflow. This will provide an opportunity for investigation of usability testing, a widely neglected topic in literature text mining. Building on the coauthors' extensive experience in evaluation of interactive systems [30], we will employ well-understood formal and informal techniques for user interface evaluation—those specific to search interfaces [31] or in general [32]—to address the lack of research into user interface design for biomedical text mining tools for curators.

5. Conclusion

We have presented a basic framework, iProLINK, to link the text mining tool developers to the ontology and systems biology user/curator communities. We used several use cases to illustrate the need and feasibility of bridging disparate communities, and analyzed requirements of the interface and major gaps in the community effort. A well designed interface and community workshops for curation and evaluation of tools will be the keys for success.

6. Acknowledgements

The work at PIR (HL, ZZH, CHW) was supported by National Science Foundation (NSF) Grant IIS-0639092, and US Army TATRC #W81XWH0720112. The work at MITRE (KBC, LH) was supported by NSF Grant II-0640153. The work of CNIO was supported by grant ENFIN NoE LSHG-CT-2005-518254.

References

[1] M.Y. Galperin. The Molecular Biology Database Collection: 2008 update. *Nucleic Acids Res.* 36(Database issue):D2-4, 2008.

[2] Z.Z. Hu, I. Mani, V. Hermoso, H. Liu and C.H. Wu. iProLINK: an integrated protein resource for literature mining. *Comput Biol Chem* 28: 409-416, 2004.

[3] C.H. Wu, L.S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, et al. The Protein Information Resource. *Nucleic Acids Research*, 31: 345-347, 2003.

[4] I. Mani, Z.Z. Hu, S.B. Jang, K. Samuel, M. Krause, J Phillips, C.H. Wu. Protein name tagging guidelines: lessons learned. *Comparative and Functional Genomics* 6:72-76. 2005.

[5] Z.Z. Hu, M. Narayanaswamy, K.E. Ravikumar, K. Vijay-Shanker, C.H. Wu. Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics* 21(11): 2759-2765, 2005.

[6] H. Liu, Z.Z. Hu, C.H. Wu. BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics* 22:103-105, 2006.

[7] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, Overview of BioCreative: Critical Assessment of Information Extraction for Biology, *BMC Bioinformatics* 6(Suppl 1):S1, 2005.

[8] M. Krallinger, A. Morgan, L. Smith, F. Leitner, Tanabe, et al. Evaluation of text mining systems for Biology: overview of the Second BioCreative community challenge, *Genome Biology*, 9(Suppl 2):S1, 2008.

[9] W. Hersh, A. Cohen, L. Ruslen, P. Roberts: TREC 2007 Genomics Track Overview. In *Proceedings of the Sixteenth Annual Text REtrieval Conference - TREC 2007;* 2007; Gaithersburg, MD.

[10] T.G. Dietterich: Ensemble methods in machine learning. *Multiple Classifier Systems*, 1857:1-15, 2000.

[11] Y.S., Chung, D.F. Hsu, C.Y. Tang: On the Diversity-Performance Relationship for Majority Voting in Classifier Ensembles. In: *7th International Workshop on Multiple Classifier Systems:* Springer Verlag; 2007.

[12] L. Si, T. Kanungo and X. Huang. Boosting Performance of Bio-Entity Recognition by Combining Results from Multiple Systems, In *Proc of Workshop on Data Mining in Bioinformatics*, 2005, pp. 76–83.

[13] J. Wilbur, L. Smith and L. Tanabe. BioCreative 2. Gene Mention Task, In *Proc of the Second BioCreative Challenge Evaluation Workshop*, 2007, pp. 7–16.

[14] L. Smith, L. Tanabe, R. Ando, C. Kuo, J. Chung, et al. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(Suppl 2):S2, 2008.

[15] F. Leitner, M. Krallinger, C. Rodriguez-Penagos, J.Hakenberg, C Plake et al. Introducing meta-services for biomedical information extraction. *Genome Biology*, 9(Suppl 2):S6, 2008.

[16] GENIA Project (2001) GPML overview. http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/GPML/.

[17] K.B. Cohen, L. Fox, P.V. Ogren, and L. Hunter. Corpus design for biomedical natural language processing. Linking Biological Literature, Ontologies and Databases, 2005, pp. 38-45. Association for Computational Linguistics.

[18] D. Ferrucci and A. Lally. Building an example application with the unstructured information management architecture. *IBM Systems Journal* 43(3):455-475, 2004.

[19] R. Mack, S. Mukherjea, A. Soffer, N. Uramoto, E. Brown, et al. Text analytics for life science using the unstructured information management architecture. *IBM Systems Journal* 43(3):490-515, 2004.

[20] W.A. Baumgartner Jr., K.B. Cohen, and L. Hunter. An open-source framework for large-scale, flexible evaluation of biomedical text mining systems. *Journal of Biomedical Discovery and Collaboration* 3(1), 2008.

[21] Y. Kano, N. Nguyen, R. Sætre, K. Yoshida, et al. Filling the gaps between tools and users: A tool comparator, using protein-protein interactions as an example. *Pacific Symposium on Biocomputing,* 6:616-627, 2008.

[22] H.L. Johnson, W.A. Baumgartner Jr., M. Krallinger, K.B. Cohen, L. Hunter. Corpus refactoring: a feasibility study. *Journal of Biomedical Discovery and Collaboration* 2(4), 2006.

[23] B. Smith, M. Ashburner, C. Rosse, J Bard, W. Bug, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 25:1251-5, 2007.

[24] D.A. Natale, C.N. Arighi, W. Barker, J. Blake, T. Chang, et al. Framework for a Protein Ontology. *BMC Bioinformatics*, 8(Suppl 9):S1, 2007

[25] H. Huang, Z.Z. Hu, C.N. Arighi, C.H. Wu. Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. *Front Biosci.* 12:5071-88, 2007.

[26] S. Kim, S.Y. Shin, I.H. Lee, S.J. Kim, R. Sriram, B.T. Zhang. PIE: an online prediction system for protein-protein interactions from text. *Nucleic Acids Res.* 36(Web Server issue):W411-5, 2008.

[27] J.M. Fernández, R. Hoffmann, A. Valencia. iHOP web services. *Nucleic Acids Res*. 35:W21-6, 2007.

[28] M.S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, et al. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc.* 2:2366-82, 2007.

[29] Z.Z. Hu, H. Huang, B. Kagan, A. Riegel, A. Wellstein, A. Dritschilo, C.H. Wu. Protein-centric integration and functional analysis of cancer omics data. *US HUPO Annual Conference* March 16-19, 2008, Washington DC.

[30] J. Polifroni, L. Hirschman, S. Seneff, and V. Zue. Experiments in evaluating interactive spoken language systems. *Human Language Technology Conference: Proceedings of the Workshop on Speech and Natural Language,* 1992, pp. 28-33.

[31] M. Hearst, "Evaluation of Search Interfaces," *Modern Information Retrieval*, 2nd Edition: The Use and Technology behind Search Engines, R. Baeza-Yates, B. Ribeiro-Neto, and M. Hearst, Addison-Wesley, to appear

[32] B. Shneiderman and C. Plaisant. *Designing the User Interface.* Addison Wesley. 2004.

PAG XVII - January 10-14, 2009
http://www.intl-pag.org/

PIR Workshop - January 11, 8:00-10:10am

Linking Text Mining with Ontology and Systems Biology for Database Curation

Zhang-Zhi Hu and Lynette Hirschman

ABSTRACT

The rapid growth of scientific literature and of large-scale molecular data has prompted active research in biological text mining to facilitate literature-based database curation. Meanwhile, systems biology knowledgebases and ontologies are emerging as critical tools in biological research where complex data in disparate resources need to be integrated and annotated. The challenge for text mining is to develop tools that will be broadly utilized by biological user communities.

1. *Literature mining resource*: PIR has developed iProLINK as a resource to support text mining and NLP research. It provides literature corpora with annotation-tagged abstracts for training and benchmarking text mining tools, as well as tools such as RLIMS-P for mining protein phosphorylation information and BioThesaurus for resolving synonyms and ambiguous names of genes and proteins.
2. *BioCreative Challenge Evaluations*: Bringing together text mining users with tool developers, there have been two BioCreative evaluations to date. The first focused on functional annotations, including linkage of evidence passages to support Gene Ontology annotations for proteins in text, while the second focused on extraction of protein-protein interaction (PPI) information, using "gold standard" data provided by PPI databases.
3. *Linking text mining with ontology and systems biology*: Built on iProLINK, PIR is developing a framework to integrate public text mining tools, focusing on interface design and biological use cases in the context of ontology and systems biology. The framework will allow biologists to mine PPI information from the scientific literature and evaluate utility and usability of the tools for database curation and knowledge discovery.

Workshop organization:
Session 1: 8:00-8:40 am, Zhang-Zhi Hu
Session 2: 8:40-9:20 am, Lynette Hirschman
Session 3: 9:20-9:50 am, Zhang-Zhi Hu
Open Discussion: 9:50-10:10am, Lynette Hirschman and Zhang-Zhi Hu

Note: The workshop will consist of 3 sessions, each including a presentation and 10 min discussion, followed by a 20-min open discussion.